

海量医疗数据跨域知识安全迁移全生命周期技术研究^{*}

陆贵强

白玉真

金博 刘一玮

(辽宁装备制造职业技术学院资源与
信息技术服务中心 沈阳 110004)

(中国信息通信研究院云计算与
大数据研究所 北京 100191)

(大连理工大学 大连 116024)

周海涛

陈如梵 李 帜

(云赛智联股份有限公司上海数翎信息技术
分公司 上海 200434)

(杭州诺威信息科技有限公司 杭州 310053)

〔摘要〕 介绍国内外医疗健康大数据应用现状, 基于现实中海量医疗数据可能存在的各类异质性表现, 依据信息论、博弈论等技术, 提出分析不同数据源对目标任务的潜在价值并据此优选数据源的方法。

〔关键词〕 医疗数据; 知识迁移; 全生命周期管理; 跨域安全

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2022.10.002

Study on Full Lifecycle Technology for Cross-domain Knowledge Security Migration of Massive Medical Data LU Guiqiang, Informatization and Resource Center, Liaoning Equipment Manufacturing Vocational-Technical College, Shenyang 110034, China; BAI Yuzhen, Cloud Computing & Big Data Research Institute, China Academy of Information and Communications Technology, Beijing 100191, China; JIN Bo, LIU Yiwei, Dalian University of Technology, Dalian 116024, China; ZHOU Haitao, INESA Intelligent Tech Inc. Shanghai Shuyi Information Technology Branch, Shanghai 200434, China; CHEN Rufan, LI Zhi, Hangzhou Nuowei Information Technology Co., Ltd, Hangzhou 310053, China

〔Abstract〕 The paper introduces the application status of medical and health big data at home and abroad. Based on the heterogeneity of massive medical data in reality, and based on information theory, game theory and other technologies, it proposes the methods to analyze the potential value of different data sources for target tasks and select data sources accordingly.

〔Keywords〕 medical data; knowledge transfer; full lifecycle management; cross-domain security

〔修回日期〕 2022-10-12

〔作者简介〕 陆贵强, 高级工程师; 通信作者: 金博, 教授, 博士生导师, 发表论文 113 篇。

〔基金项目〕 国家自然科学基金资助项目“面向创新治疗方案推荐的药物可重用路径挖掘方法”(项目编号: 62172074)。

1 引言

近年来人工智能(Artificial Intelligence, AI)的飞速发展在引领新一轮全球技术革命的同时也暴露出其在隐私保护和数据安全等方面所面临的严峻挑战。有研究表明通过对 AI 系统的攻击可以推理

出训练数据中的个体敏感信息,或者恶意篡改机器学习结果。因此发展高效、可信、鲁棒的安全人工智能基础理论及技术已经成为国内外共同关注的问题,建设新一代人工智能已成为国家战略。在医疗健康领域,涉及海量患者个人隐私数据,如何在确保安全的条件下对数据进行迁移,并实现全生命周期的数据管理至关重要。本文针对非均匀医疗数据源,研究最小化信息交换的数据源迁移有效性度量方法;针对数据源异质安全需求,构建融合差分隐私、多方可信计算的层次化跨域知识安全汇聚机制;针对时变演化的数据源,研究知识汇聚模型的模块化适时更新机制。为实现面向健康医疗数据的全生命周期管理,海量数据源的知识汇聚和迁移将面临计算效率、数据安全、时变演化的挑战。本文重点研究隐私保护下海量医疗数据源高效率、强安全、自适应的知识迁移汇聚技术。针对海量数据源,研究最小化信息交换的数据源迁移价值度量方法,实现高效准确的迁移质量评估和高价值数据源优选;根据数据源异质安全需求,构建融合差分隐私、多方可信计算、同态加密等多种安全机制的层次化跨域知识安全汇聚方法,实现安全、效率双优化;面对时变数据源,研究知识汇聚模型的自适应适时演化机制,实现质效优化下的汇聚模型动态生长。基于上述研究成果,本文提出一套隐私保护下海量数据源跨域知识安全迁移全生命周期技术框架,支持现有开源机器学习平台和海量医疗数据。

2 国内外应用现状

2.1 国内

随着信息技术不断发展,大部分医院建立了各类信息管理系统,特别是伴随物联网、移动医疗、自动化分析检测仪、可穿戴设备的普及,医院、医生和患者都成为数据的直接创造者,每天产生海量医疗健康数据^[1-2]。充分挖掘医疗健康大数据的潜在价值,对提高医疗质量、节约医疗成本、加强个人健康管理有重要作用^[3-4]。在我国,医疗健康大数据也越来越受到重视。《“健康中国 2030”规划纲要》中明确指出加强健康医疗大数据应用体系建设,推进基于区域人口健康信息平台的医疗健康大

数据开放共享、深度挖掘和广泛应用。健康医疗大数据可为医务工作者、患者提供医疗知识,在恰当的时间智能化地过滤和表达信息,使临床医疗达到最佳疗效,具有极其重要的意义^[5]。而治疗用药决策作为临床决策的重要组成部分,其可根据历史医疗健康大数据来辅助医生更加高效地选择和制定有益于患者的最佳治疗方案和用药组合,从而更好地缓解医疗资源欠缺的现状。医疗大数据具有数据量大 (Volume)、实时性强 (Velocity)、种类多样 (Variety) 和潜在价值高 (Value) 4 个特点^[6],潜在价值挖掘面临巨大挑战。例如临床数据和实验室数据整合在一起,数据量巨大,仅靠人工经验完全无法识别出真正具有高价值的病历参考。医疗健康大数据中包含着大量的多元异构以及多模态数据,如电子病历数据中包含患者个人信息、历史用药数据、历史检查检验等结构化数据,病情描述等文本类型数据,CT 等图像类数据,且此类数据具有一定的相关性和时序复杂性。合理高效地通过智能决策方法对这些医疗数据进行分析处理,既能够使医生诊疗有迹可循,还可以发现最有效的治疗方案或者用药方法,从而及时为医生和患者提供最佳的诊疗建议^[7]。

2.2 国外

现有开源联邦学习框架^[8]主要基于 2016 年提出的经典联邦学习范式^[9],该范式假设各数据源同质,可采用同一联邦学习算法汇聚知识,并共享单一联邦学习模型。但在机器学习实践中,数据源广泛存在异质性,包括数据分布、样本覆盖、特征维度、优化目标、隐私需求等,往往难以满足同质性假设。一些国内外高校和机构针对异质性改进现有联邦学习框架,提出个性化联邦学习、联邦迁移学习等新型隐私机器学习范式^[10],但仍主要集中在解决数据非独立同分布、样本覆盖不一致等有限异质性问题。拓展现有联邦学习框架,突破海量数据源间存在的各类异质性将有望扩展隐私机器学习的实践及应用范围。

3 医疗数据全生命周期跨域迁移

3.1 最小化信息交换的数据源迁移价值度量优选方法和技术

在给定医疗数据迁移场景中,选择合适数据源提

取知识,是成功完成知识迁移的先决条件。在面对海量数据源并满足隐私保护的要求下,除确保数据源价值度量准确性外,还需兼顾高效性与安全性两个核心指标。本文在现有基于博弈论的公平数据价值度量方案基础上,探索如何最小化价值度量流程中数据源间所需交换的信息量,以及如何高效选择最优数据源进行知识迁移,同时达到降低隐私泄露风险和提升计算效率的双重目标。特别针对不同数据源组合可能展现出的不同特性,如是否满足随数据源增加、模型边际效应递减属性等,分别研究基于运筹优化理论(如子模优化)和机器学习梯度优化理论的数据源优选算法,达到质效双优,见图1。

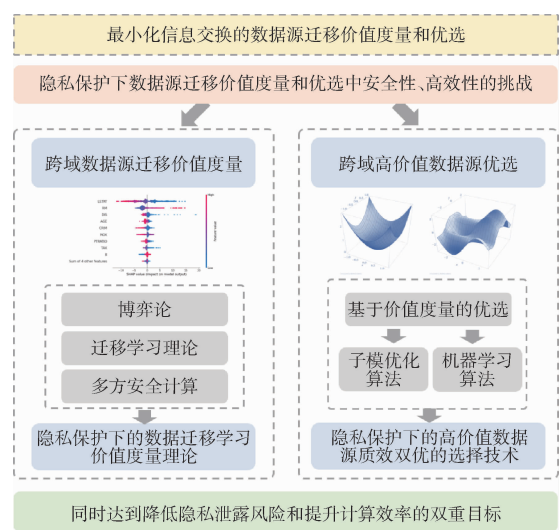


图1 最小化信息交换的数据源迁移价值度量优选方法和技术

3.2 基于异构安全机制的层次化跨域知识汇聚方法和技术

异构安全机制(如差分隐私、多方安全计算、同态加密等)可达到不同保护效果,适配数据源不同场景下(例如医疗数据在临床诊断和医保核算等场景)的个性化隐私需求。传统“服务器-数据源”单层知识汇聚范式难以适配数据源异质隐私需求,且服务器直接与海量数据源对接,计算通信效率难以优化。本文突破传统的知识汇聚单层架构,探索基于异构安全机制的层次化跨域知识汇聚新范式。研究如何设计面向海量数据源个性化隐私需求的层次化组织和知识流通架构,如何实现针对各层

异质隐私需求的知识安全汇聚机制,以构建质效最优、可灵活扩展的知识安全汇聚方法,见图2。

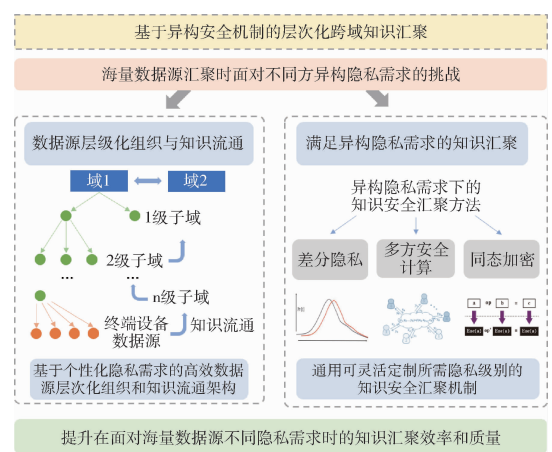


图2 基于异构安全机制的层次化跨域知识汇聚方法和技术

3.3 动态自适应的汇聚模型适时演化方法和技术

在医疗环境中医疗数据每时每刻不断产生,而数据源的时变将影响其在汇聚模型中的迁移效果;既往有效迁移源可能逐渐失效,新兴有效迁移源则持续产生。因此,动态演化汇聚模型才能长期维持高质量的知识迁移。本文研究如何基于实时流数据处理技术,针对海量数据源安全、高效、快速地检测其迁移价值的时变规律,发现潜在的价值突变;特别是在数据源无法保持稳定连接时准确检测其价值变化。进一步针对不同价值变化的数据源设计自适应的汇聚模型适时演化算法,删除价值显著降低的数据源,新增具备显著价值且未纳入的数据源,以及更新价值较高的已有数据源,见图3。

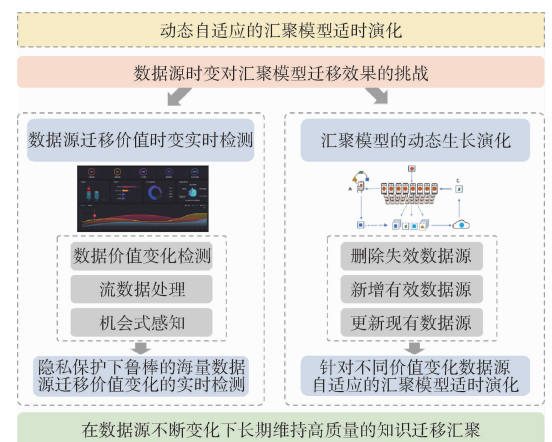


图3 动态自适应的汇聚模型适时演化方法和技术

4 实验验证

4.1 实验目的

形成一整套隐私保护下海量数据源跨域知识迁移汇聚全生命周期的工具库，包含迁移数据源的价值度和优选，基于异构安全机制的知识迁移汇聚以及汇聚模型的自适应更新演化等功能，支持开发人员快速建立和部署质效优化且安全的跨域知识迁移汇聚系统。该工具库将通过对主流开源机器学习框架所提供功能接口的进一步抽象，实现无缝衔接各主流框架，支持海量数据源知识汇聚，基于该工具库开展实验验证。根据国家对医保支付方式改革的要求，需要通过优化病案首页管理以及患者住院费用（包括药品费用、耗材费用等）数据的智能化分析，建立医保费用动态测算模型，加强和升级地区医保费用的稽核控费；强化地区医疗机构横向比较和评估管理，做到同级别医疗机构同病、同治、同质、同价，并为全面开展按疾病诊断相关分组（Diagnosis Related Groups, DRG）收付费管理进行前期准备。

4.2 具体任务

按病案首页模式对医保与医疗机构的结算接口进行改造；本地医疗机构病案首页明细、住院费用明细、药品清单、耗材清单等数据采集及清洗；药品统一编码和赋值；出院患者的药品数据采集、清洗、分析；耗材统一编码和赋值；出院患者耗材数据采集、清洗、分析；构建地区内住院患者费用明细视图、病案首页信息视图，建立地区疾病谱；根据地区内参保类别、医疗机构等级进行报销比例设置和分类别费用总额管理；特殊药物（抗菌药物、毒麻药品等）的多维度精细化管理；特殊高值药品和耗材的多维度精细化管理；与地方疾病谱相对的地方常用药品库、总额预付费、医保及其他各类非医保费用比例可视化管理；地区性监测指标的自定义管理；考虑地方实际情况，对特殊支付政策制定进行数据支持；极端患者另行支付、加大审核；死亡和转诊问题；特殊高值耗材如何打包问题；地方常见疾病费用结算、费率调整（如呼吸系统疾

病、心血管系统疾病等）。这是本研究团队为某三甲医院实施部署的医疗服务价格监管系统，本文提出的知识安全迁移技术应用在该项目中，实现了在医保管理机构监督下的数据迁移操作，具体任务是从临床诊疗场景向医保核算场景进行数据迁移，涉及需要进行安全迁移的任务内容，见图 4。

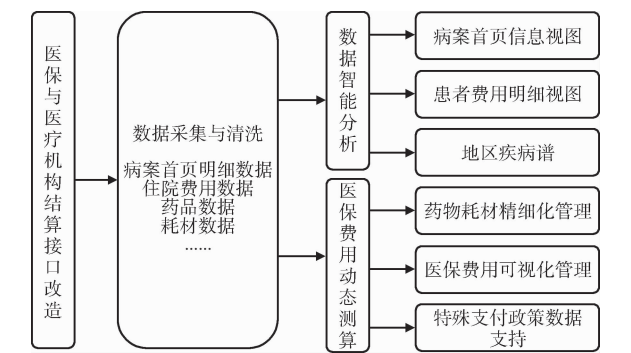


图 4 医疗服务价格管理系统数据处理与迁移任务

4.3 实验结果

基于本文提出的技术，在海量数据迁移需求调节下，项目团队顺利完成并行数据处理，数据迁移性能测试结果，见表 1。经过数据完整性测试以及大量有效和无效数据的测试，本文提出的技术对于有效数据可以正常运行，对于无效数据，可以对异常以及非法输入进行处理，不会给用户带来损失。根据测试数据分析可以看出系统平均响应时间均在 1 秒以内，并发数 20 ~ 200 不等，与传统数据迁移方向相比，在保证数据安全的前提下，提升数据迁移效率 20% 以上。经对比，在保证数据安全和高效迁移的条件下，数据迁移前后数据分布无变化，达到预期目标，完成并通过项目验收。

表 1 数据迁移性能测试

模块	并发数	平均响应时间（毫秒）
项目管理模块	20	832
耗材管理模块	25	133
项目耗材关联模块	25	81
预测管理模块	100	409
项目对比模块	80	415
项目分类管理	200	331
项目阈值管理	200	351
耗材类别管理	200	348

（下转第 29 页）

uance: an Expectation - confirmation Model [J]. *Mis Quarterly*, 2001, 25 (3): 351 - 370.

15 Fazio R H, Sanbonmatsu D M, Powell M C, et al. On the Automatic Activation of Attitudes [J]. *Journal of personality and Social Psychology*, 1986, 50 (2): 229.

16 魏文冰. Web2.0 环境下虚拟学习社区归属感研究 [D]. 重庆: 西南大学, 2014.

17 Zhao L, Lu Y B, Wang B, et al. Cultivating the Sense of Belonging and Motivating User Participation in Virtual Communities: a Social Capital Perspective [J]. *International Journal of Information Management*, 2012, 32 (6): 574 - 588.

18 Wixom B H, Todd P A. Theoretical Integration of User Satisfaction and Technology Acceptance [J]. *Information Systems Research*, 2005, 16 (1): 85 - 102.

19 Lederman R, Fan H M, Smith S, et al. Who Can You Trust? Credibility Assessment in Online Health Forums [J]. *Health Policy and Technology*, 2014, 3 (1): 13 - 25.

20 Davis F D. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology [J]. *Mis Quarterly*, 1989, 13 (3): 319 - 340.

21 张沪月, 许统, 熊芳. 用户多维度阅读交互对阅读满意度的影响效应研究 [J]. *图书情报知识*, 2017 (3):

29 - 37.

22 Lin H F. Understanding Behavioral Intention to Participate in Virtual Communities [J]. *Cyber Psychology & Behavior*, 2006, 9 (5): 540 - 547.

23 Koh J, Kim Y G. Sense of Virtual Community: a Conceptual Framework and Empirical Validation [J]. *International Journal of Electronic Commerce*, 2003 (8): 75 - 93.

24 王文韬, 谢阳群, 刘坤锋. 基于扎根理论的虚拟健康社区用户使用意愿研究 [J]. *情报资料工作*, 2017 (3): 75 - 82.

25 董庆兴, 周欣, 毛凤华, 等. 在线健康社区用户持续使用意愿研究——基于感知价值理论 [J]. *现代情报*, 2019 (3): 3 - 14.

26 王哲. 社会化问答社区知乎的用户持续使用行为影响因素研究 [J]. *情报科学*, 2017, 35 (1): 78 - 83.

27 王国华, 刘菊, 杨腾飞, 等. 网络空间中艾滋病的社会支持研究——以百度贴吧“HIV 吧”为例 [J]. *情报杂志*, 2015, 34 (11): 105 - 110.

28 宋慧玲, 帅传敏, 李文静. 知识问答社区用户持续使用意愿的实证研究 [J]. *信息资源管理学报*, 2019, 9 (4): 68 - 81.

(上接第 11 页)

5 结语

当前医疗场景下各数据源存在广泛异质性且互不可见, 为知识迁移带来巨大挑战。本文提出“来源价值优选 - 众源层次汇聚 - 适源动态演化”3 阶段的隐私保护下海量数据源跨域知识迁移的全生命周期创新性技术框架。提出安全跨域迁移价值度量方案, 开发最优迁移源选择算法; 构建层次化数据源组织架构, 建立可适配不同数据源层的异构安全技术跨域知识迁移汇聚方案; 设计隐私保护下数据源时变实时监测算法, 对知识汇聚模型进行动态演化。以上框架可以实现安全的数据迁移和全生命周期医疗数据有效管理。

参考文献

1 张振, 周毅, 杜守洪, 等. 医疗大数据及其面临的机遇与挑战 [J]. *医学信息学杂志*, 2014, 35 (6): 2 - 8.

2 Barrett M A, Humblet O, Niatt R A, et al. Big Data and Disease Prevention: from Quantified Self to Quantified Com-

munities [J]. *Big Data*, 2013, 1 (3): 168 - 175.

3 Cios K J, Moore G W. Uniqueness of Medical Data Mining [J]. *Artificial Intelligence in Medicine*, 2002, 26 (1 - 2): 1 - 24.

4 周光华, 辛英, 张雅洁, 等. 医疗卫生领域大数据应用探讨 [J]. *中国卫生信息管理杂志*, 2013 (4): 296 - 300.

5 Ghazavi S N, Liao T W. Medical Data Mining by Fuzzy Modeling with Selected Features [J]. *Artificial Intelligence in Medicine*, 2008, 43 (3): 195 - 206.

6 张大庆. 临床决策: 医学哲学研究的一个重要领域 [J]. *医学与哲学*, 2004, 25 (12): 17 - 20.

7 颜延, 秦兴彬, 樊建平, 等. 医疗健康大数据研究综述 [J]. *科研信息化技术与应用*, 2014, 5 (6): 3 - 16.

8 Schneider P J. Pharmacists Building a Safer Health System [J]. *American Journal of Health - System Pharmacy*, 2001, 58 (1): 66 - 68.

9 周传鑫, 孙奕, 汪德刚, 等. 联邦学习研究综述 [J]. *网络与信息安全学报*, 2021, 7 (5): 77 - 92.

10 吴佳慧. 机器学习中隐私保护数据计算研究 [D]. 重庆: 西南大学, 2021.